MIMIC-CXR-VQA: A Medical Visual Question Answering Dataset Constructed with LLaMA-based Annotations

Mohamed Aas-Alas[*1] maasala@prhlt.upv.es
Miquel Obrador-Reina[*1] mobrrei@prhlt.upv.es
Luis-Jesus Marhuenda[*1] ljmarten@prhlt.upv.es
Alberto Albiol[1] alalbiol@prhlt.upv.es
Roberto Paredes[1] rparedes@prhlt.upv.es
[1] *Campus de Vera, Universitat Politècnica València, Camí de Vera s/n, 46022 Valencia, Spain*

Editors: Under Review for MIDL 2026

Abstract

The interpretation of chest X-rays (CXRs) is a critical yet time-consuming task in clinical radiology, often limited by the availability of expert radiologists. To address this challenge, we introduce a new large-scale medical Visual Question Answering (VQA) dataset derived from the MIMIC-CXR database, containing over 3.2 million question-answer (QA) pairs across 15 clinically relevant categories, enabling the training of multimodal models capable of answering a broad range of diagnostic questions directly from chest X-ray images. Unlike prior datasets, our QA pairs are generated using a large language model, LLaMA 3.1, guided by a carefully crafted prompt structure to produce rich, nuanced, and evidence-based textual answers grounded in radiology reports. We address limitations of existing datasets such as templated responses and linguistic monotony by ensuring diversity, completeness, and clinical fidelity in our QA pairs. To support benchmarking on this new dataset, we provide initial baseline models and training strategies designed to evaluate visual and textual reasoning performance in the medical VQA setting. Extensive experiments demonstrate the effectiveness of our approach across multiple evaluation metrics, establishing a strong benchmark for future research in medical VQA. Our dataset and baseline models pave the way for building clinically meaningful AI tools that can assist radiologists by answering complex diagnostic questions with accuracy and interpretability.

Keywords: Visual Question Answering, Medical Imaging, Datasets.

1. Introduction

The rapid advancement of medical imaging technologies has significantly enhanced the capabilities of healthcare systems, particularly in radiology. Chest X-rays (CXRs) are among the most commonly used diagnostic tools, providing critical insights into various thoracic conditions. However, the interpretation of these images often requires specialized expertise, which can be a bottleneck in healthcare delivery. One of the primary challenges in radiology is the generation of accurate and informative radiology reports. Radiology report generation aims to convert medical images into descriptive and clinically meaningful textual interpretations. Despite advancements in deep learning, automatic report generation

[*] Contributed equally

remains challenging due to the complexity of medical terminology, variability in reporting styles, and the need for high clinical accuracy.

To address these challenges, we introduce a new, large-scale medical Visual Question Answering (VQA) dataset that is more consistent with radiologists' practice. VQA systems, which combine computer vision and natural language processing (NLP), enable automated systems to answer natural language questions about visual content, making them highly applicable to medical imaging. By focusing on identifying findings in medical images and answering clinically relevant questions, our approach aligns more closely with radiologists' diagnostic workflow.

In this study, we present a novel VQA dataset derived from the MIMIC-CXR dataset (Johnson et al., 2019), a large-scale public database of chest radiographs containing 227,835 studies, each with a unique radiology report and corresponding images. The dataset includes a total of 377,110 images. Our dataset is designed to facilitate the development of VQA models capable of answering clinically relevant questions about chest X-rays. It includes 15 categories of questions: Atelectasis, Cardiomegaly, Consolidation, Edema, Enlarged Mediastinum, Fracture, Lung Lesion, Lung Opacity, No Finding, Pleural Effusion, Pleural Other, Pneumonia, Pneumothorax, Support Devices, and Heart-related conditions. The dataset contains a total of 3,247,512 question-answer (QA) pairs, making it one of the most comprehensive resource for medical VQA research.

A key aspect of our dataset creation process is leveraging a large language model (LLM) to generate answers with the objective of producing a rich and diverse dataset. Unlike traditional approaches where VQA tasks are often reduced to simple classification problems with systematic and templated responses, our dataset ensures that models trained on it learn to understand and generate language effectively. The inclusion of detailed and nuanced answers allows models to develop deeper linguistic and clinical reasoning skills, making them more useful in real world applications.

To evaluate the usefulness of the dataset, a set of baseline models were implemented to assess performance under different training strategies.

Our contributions are summarized as follows:

- A new medical image VQA task designed to mirror radiologists' diagnostic workflow by generating clinically grounded questions and rich answers that reflect real interpretive steps rather than simple label prediction.

- A new large-scale dataset from MIMIC-CXR containing 3,247,512 LLM-generated QA pairs across 15 clinically relevant categories, explicitly avoiding systematic classification-like answers to encourage rich, diverse, and naturally expressed responses.

- Two baselines and a benchmark to guide future research toward more effective and accessible medical VQA systems.

We release both the dataset and the baseline codebase [1] publicly to encourage further developments in clinically meaningful AI systems.

[1] https://github.com/LightVED-prhlt/MIMIC-CXR-VQA-Dataset_Creation

2. Related work

Medical Visual Question Answering (VQA) is an emerging field that integrates computer vision and natural language processing (NLP) to develop systems capable of answering clinical questions based on medical images. These questions can be open-ended, requiring free-text answers that involve reasoning, or closed-ended, which typically expect a limited set of predefined responses such as yes/no or single-word choices. Several datasets have been introduced to facilitate research in medical VQA, each with unique characteristics and challenges.

One of the pioneering medical VQA datasets is VQA-RAD (Lau et al., 2018), which contains 3,515 QA pairs generated by clinicians and 315 radiology images covering the head, chest, and abdomen. Each image is associated with multiple questions, categorized into 11 distinct types, including abnormality, modality, organ system, size, positional reasoning, and others. The dataset includes a balanced mix of binary answers (yes/no) and short phrases or single-word responses.

A more recent addition to medical VQA datasets is SLAKE (Liu et al., 2021a), a Semantically-Labeled Knowledge-Enhanced dataset designed for radiology-based VQA. The dataset consists of 642 radiology images and over 7,000 diverse QA pairs annotated by experienced physicians. Unlike earlier datasets, SLAKE incorporates external medical knowledge through a structured knowledge graph, enriching the reasoning capabilities required for answering complex questions. Additionally, SLAKE provides extensive visual annotations, including semantic segmentation masks and object detection bounding boxes. It also covers a broader range of anatomical regions, including the brain, neck, chest, abdomen, and pelvic cavity. Notably, SLAKE is a bilingual dataset, available in both English and Chinese.

A significant advancement in medical VQA datasets came with the release of PathVQA (He et al., 2020), which specifically focuses on pathology images. PathVQA comprises 4,998 pathology images and 32,799 QA pairs. Each image is associated with multiple questions related to various aspects such as location, shape, color, and appearance. This dataset is particularly valuable for developing models tailored to microscopic medical imaging, enabling deeper understanding and reasoning within the domain of histopathology.

Table 1: Comparative Analysis Across Different Medical VQA Datasets

Dataset	VQA-RAD		SLAKE			PathVQA			Medical-Diff-VQA			MIMIC-CXR-VQA(**Ours**)		
	Train	Test	Train	Val	Test	Train	Val	Test	Train	Val	Test	Train	Val	Test
Images	313	203	450	96	96	2599	858	858	131825	16402	16431	249006	2098	3830
QA Pairs	1797	451	4919	1053	1061	19755	6279	6761	560563	70070	70070	3157795	26627	63090
-Open	770	179	2976	631	645	9949	3144	3370	187920	23348	23592	3151643	26586	62952
-Closed	1027	272	1943	422	416	9806	3135	3391	372643	46722	46478	6148	41	138

However, these datasets are relatively small and primarily used for testing purposes rather than training robust medical VQA models. Larger datasets are required to train deep learning models effectively. One such dataset is Medical-Diff-VQA (Lu et al., 2024), which significantly expands the number of QA pairs compared to the first two mentioned. Medical-Diff-VQA contains a much larger number of question-answer pairs, providing better coverage of medical knowledge. However, despite its size, it has notable limitations. The dataset was generated based on predefined rules, which means that the answers follow a rigid and repetitive structure. As a result, it lacks linguistic richness and variability. Furthermore, since the answers essentially enumerate the findings relevant to the question without forming natural sentences, the dataset can be reduced to a classification problem rather than a true generative task. This issue limits its effectiveness in training models for real-world applications where complex sentence structures and nuanced medical reasoning are required.

To address these shortcomings, we have created a large-scale medical VQA dataset that overcomes the limitations of existing datasets. As shown in Table 1, our dataset contains a significantly larger number of diverse and contextually rich QA pairs, generated by leveraging a LLM. Unlike previous datasets with templated responses, our approach introduces variability in linguistic expression and clinical reasoning, and because almost all answers in our dataset are open-ended, it is considerably more suitable for real-world clinical scenarios where free-text reasoning is required. This enables models to develop deeper linguistic comprehension, ultimately improving their applicability in practical radiological and medical settings.

3. Dataset Creation

The new VQA dataset was generated from the MIMIC-CXR dataset (Johnson et al., 2019). To achieve this, we utilized a LLM and a structured set of steps to generate QA pairs from each medical report. The QA pairs were created by leveraging CheXpert labels (Irvin et al., 2019) and predefined question designed to extract meaningful information from the reports. The rest of this section describes the question generation process and the automatic answer generation pipeline.

3.1. Question Generation

In order to generate the questions, we utilized the CheXpert labels from the MIMIC-CXR dataset, which categorizes each radiology report into 14 distinct categories. Each label in this dataset can take one of four values: 1.0 (positive for the condition), 0.0 (explicitly negative for the condition), -1.0 (uncertain), or NaN (not mentioned). To construct our VQA dataset, we designed six questions for each one of these labels. The full list of questions can be found in Appendix A.

In addition to the CheXpert labels, we introduced an extra label, *heart*, after observing that the term appeared frequently in reports and carried clinical significance. If a report contained the word *heart*, we assigned a value of 1.0 to this label and ensured that the predefined heart-related questions were included.

Table 2: Distribution of CheXpert label values and corresponding generated questions.

CheXpert Labels			**Generated Questions**
Value	Count	Meaning	Count
1.0	423,065	Positive Finding	1,585,206
0.0	158,614	Negative Finding	365,424
NaN	2,532,881		1,055,232
-1.0	75,018	Uncertain Finding	241,650

Table 2 illustrates the relationship between the original CheXpert label values and the resulting number of generated questions in our dataset. The distribution reveals an imbalance among the CheXpert labels, with positive (1.0) values occurring more frequently than negative (0.0) or uncertain (-1.0) ones. To mitigate this issue and balance the frequency of questions related to negative findings, additional questions derived from labels marked as NaN were added, since any finding not mentioned in the report is considered negative.

The effectiveness of this balancing strategy is reflected in the final question distribution shown in Figure 1, where it can be seen that the VQA dataset is doubly balanced, both in terms of question coverage across labels and in the proportion of positive and negative findings.

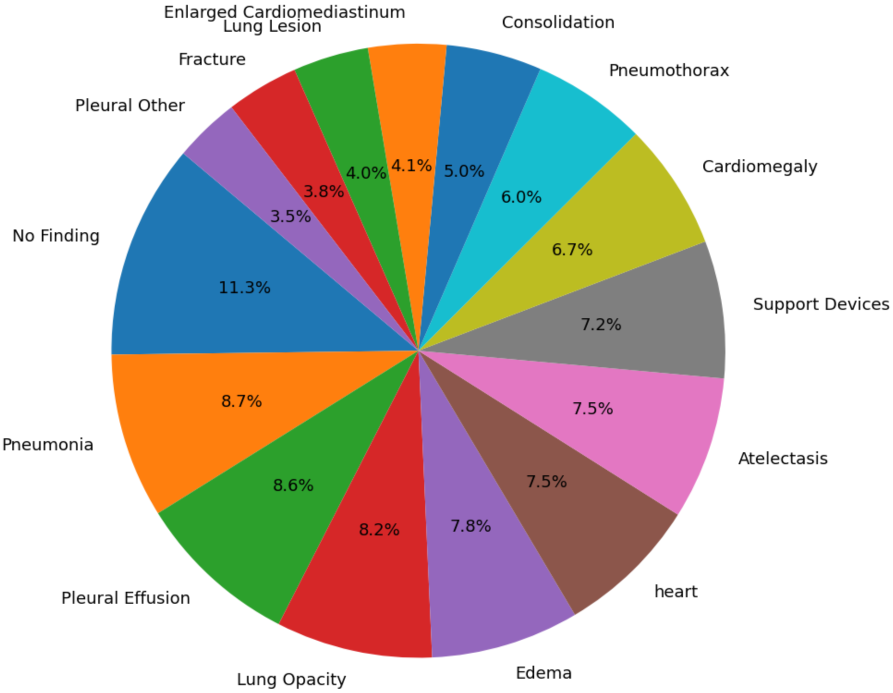

Figure 1: Final distribution of generated questions across all labels.

3.2. Automatic LLM Answer Generation

Using the previous generated questions and the reports, a large language model was used to generate the corresponding answers automatically. This was acomplished using the prompt described in Appendix B. The LLM model received the report text together with the corresponding question, ensuring that the generated answer was clinically consistent, coherent, and aligned with the radiologist's findings. It is important to mention, that additionally to the findings report, the prompt was enriched with the indications section of the report, to provide clinically relevant context as seen in Appendix C.

As explained in Appendix B, the system prompt was built incrementally by adding multiple sub-prompts. We enforced a strict response format and an evidence-based style to obtain detailed, reliable answers and prevent hallucinations by ensuring that all information originated from the report. Because radiology reports often mention comparisons with prior studies, we added a Prohibited Terms and Constraints section to eliminate such references, as our dataset targets single-study VQA without historical comparison.

Early tests revealed that some answers referenced the report itself (e.g., as written in the report), whereas we required references to the radiograph. To correct this, we introduced Terminology Guidelines to enforce consistent radiology-focused phrasing. Residual comparison-related expressions were then removed through a Neutrality and Objectivity sub-prompt, directing the LLM to answer as if observing the radiograph for the first time and ensuring purely descriptive, non-comparative outputs.

Several LLM models were considered to generate the answers for their ability to follow our prompt guidelines. For instance, we experimented with several medical fine-tuned language models, such as Meditron (Chen et al., 2023). However, after conducting multiple tests, we decided to use LLaMA (Touvron et al., 2023), as it performed exceptionally well with our prompt structure and constraints. In particular, LLaMA shined on strong language understanding and with less hallucinations than models finetuned on medical data, which aligns with our objective of generating strictly evidence-based answers without assumptions.

Among the family of LLaMA models, we used LLaMA 3.1 with 8B parameters in the Q4_K_M quantization format. This configuration provided a balance between performance and efficiency, ensuring that our system could generate high-quality, unbiased responses while maintaining computational feasibility.

The response generation process was performed separately for the training, validation, and test sets to maintain clear dataset partitions and prevent potential data leakage. Generating responses for the entire dataset required a total of eight days of processing on an RTX 4090, underscoring the computational demands of producing high-quality answers across all samples.

4. Dataset Validation

Given the critical nature of clinical QA tasks, it is essential to rigorously assess the consistency and reliability of the generated answers. The evaluation must consider four key metrics: correctness, alignment with the original report, answer completeness, and clinical relevance to ensure that the dataset is correct and maintains high fidelity to the underlying medical records.

To systematically validate the generated QA pairs, we employed a hybrid assessment strategy combining human and automated verification, as the large size of the dataset makes a full human review infeasible. Specifically, 100 QA pairs were evaluated by three human reviewers to establish a high-quality benchmark. A fourth verifier, an automated LLM-based assessor, also reviewed the same 100 pairs as an experiment to assess how closely its judgments align with those of human reviewers. After this initial validation, the LLM-based assessor evaluated a larger sample of 33,130 QA pairs derived from a randomly sampled subset of 1,000 radiology reports.

For the automated evaluation, we used DeepSeek-R1 (DeepSeek-AI et al., 2025) as the LLM, selected for its reasoning capabilities. To that end, we designed a new structured prompt instructing the model to evaluate each QA pair across the four key dimensions. This evaluation framework allows to reliably detect discrepancies between generated answers and radiology reports, supporting the dataset's suitability for downstream tasks.

Table 3: Human and LLM-based evaluation of radiology QA pairs in two scales and inter-rater agreement across verifiers.

	100 QA Pairs Evaluation				**Large-Scale Evaluation**		**Inter-Rater Agreement**		
	Human			**LLM**					
	V1	**V2**	**V3**	**V4**			κ	AC1	AC2
					Count	**%**			
Correctness									
Correct	92.0%	98.0%	96.0%	99.0%	31317	94.53%	0.2843	0.9443	0.9443
Incorrect	8.0%	2.0%	4.0%	1.0%	1813	5.47%			
Consistency with Report									
Fully Consistent	90.0%	91.0%	71.0%	92.0%	30136	90.96%	0.3291	0.7417	0.8953
Partially Consistent	8.0%	7.0%	27.0%	7.0%	1866	5.63%			
Inconsistent	2.0%	2.0%	2.0%	1.0%	1128	3.41%			
Answer Completeness									
Complete	70.0%	88.0%	88.0%	83.0%	24653	74.41%	0.3720	0.6851	0.8586
Partially Complete	26.0%	10.0%	9.0%	13.0%	5593	16.88%			
Incomplete	4.0%	2.0%	3.0%	4.0%	2884	8.71%			
Clinical Relevance									
Essential	90.0%	95.0%	95.0%	81.0%	27683	83.56%	0.3395	0.8312	0.8858
Non-essential but Correct	3.0%	3.0%	1.0%	17.0%	3410	10.29%			
Incorrect and Misleading	7.0%	2.0%	4.0%	2.0%	2037	6.15%			

As shown in Table 3, the four verifiers exhibit highly consistent judgments. We additionally computed inter-annotator agreement using weighted Fleiss' κ (Fleiss, 1971), Gwet's AC1 (Gwet, 2002), and AC2 (Gwet, 2012). Following Landis and Koch scale (Landis and Koch, 1977), Fleiss' κ indicated fair to moderate agreement, whereas AC1 and AC2 showed substantial to almost perfect agreement. This discrepancy arises because κ subtracts chance agreement, which becomes inflated when annotators overwhelmingly select the same category (e.g., labeling correctness as correct). Consequently, κ underestimates agreement despite near-unanimous evaluations. Overall, all metrics demonstrate strong alignment among verifiers, reinforcing the dataset's reliability.

As a final remark, the results in Table 3 demostrate that all verifiers converge on a consistently high correctness of the question–answer pairs, confirming that the dataset's annotations are not only reliable but also accurate in their clinical and linguistic content. This strong consensus ensures that the dataset provides a trustworthy supervisory signal, making it well suited for training linguistically rich, diverse, and clinically robust VQA models.

5. Baseline Models

We present the baseline models used to establish reference performance on our dataset, incorporating two different models and training strategies.

5.1. SwinVED-SCST

SwinVED (Marhuenda et al., 2025), is a Vision Encoder Decoder architecture composed of a Swin (Liu et al., 2021b) Transformer encoder and a 3-layer Transformer-based text decoder.

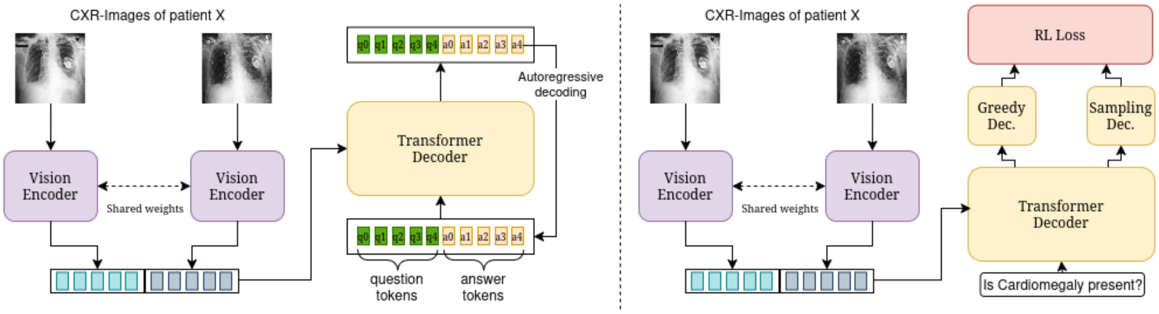

Figure 2: Architecture of our SwinDecoder-SCST baseline for MIMIC-CXR-VQA.

Our training strategy follows a three-stage process. The model architecture and the first two stages of training are directly adopted from our previous work (Marhuenda et al., 2025), where the vision encoder and language decoder are pre-trained and subsequently aligned for image–text generation tasks. In this work, we extend that pipeline by introducing a third stage.

In the Third stage (right side of Fig 2), we apply Reinforcement Learning (RL) using the Self-Critical Sequence Training (SCST) algorithm (Rennie et al., 2016). This stage aims to directly optimize a semantic similarity metric by using it as a reward signal. SCST performs two decoding passes: one using greedy decoding to generate a baseline answer Y_g, and another using multinomial sampling to produce a sampled answer Y_s, from which gradients are computed.

We utilize BERTScore (Zhang et al., 2020) as the reward function to improve the fluency and semantic accuracy of generated answers. For each generated answer, a reward is calculated by comparing Y_s and Y_g to the ground-truth reference answer Y_{ref}, and computing the difference in BERTScore values. These differences are scaled by the log-probability of the sampled answer and used as the reinforcement signal. The BERTScore-based loss is given by:

$$\text{Loss}_{\text{BERTScore}}(Y_s, Y_g, Y_{\text{ref}}) = -\left(r_{\text{BERTScore}}(Y_s, Y_{\text{ref}}) - r_{\text{BERTScore}}(Y_g, Y_{\text{ref}})\right) \log \Pr(Y_s) \quad (1)$$

The final RL loss combines this metric-specific loss with the standard negative log-likelihood loss (Loss_{NLL}) used during supervised training:

$$\text{Loss}_{\text{RL}} = \alpha \, \text{Loss}_{\text{BERTScore}} + \gamma \, \text{Loss}_{\text{NLL}} \quad (2)$$

In our experiments, we set $\alpha = 0.6$ and $\gamma = 0.4$ to balance semantic alignment with the reference text and grammatical fluency. This stage enables the model to better align its outputs with clinically and semantically meaningful language.

5.2. BLIP-2-MultiView

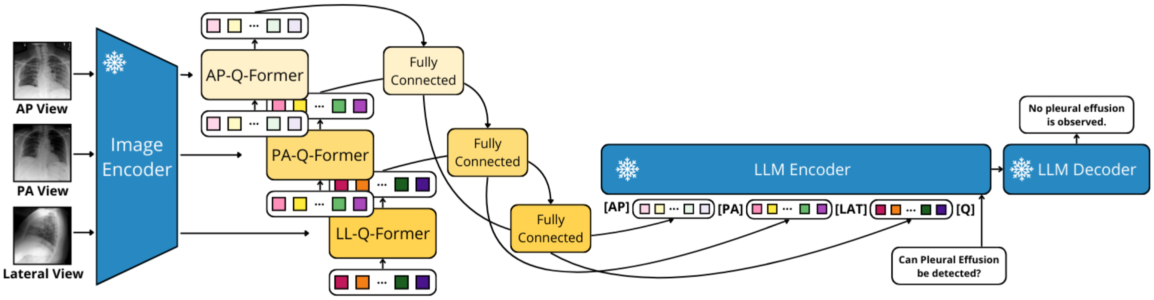

Figure 3: Architecture of the BLIP-2-MultiView baseline for MIMIC-CXR-VQA.

Baseline 2 builds on the BLIP-2 (Li et al., 2023) framework and adapts it to the multi-view nature of chest radiography. Following the BLIP-2 training strategy, both the vision encoder and the FLAN-T5 language model are kept frozen during training, and only the Q-Formers and projection layers are updated. The model employs a frozen EVA-CLIP ViT-G/14 (Sun et al., 2023) vision encoder to extract high-level visual features for each available radiographic view (anteroposterior, posteroanterior, and lateral). For every view, a dedicated Q-Former branch processes the encoder features by attending over a set of learned query tokens. Each Q-Former is paired with its own language-projection layer, which maps the 768-dimensional Q-Former outputs into the 2048-dimensional embedding space of the FLAN-T5-XL (Chung et al., 2022) language model. These projected features form three independent visual embeddings, one per view, which are subsequently concatenated into a unified visual prefix.

To preserve view-specific information, we insert a learned special token before each visual block: [AP], [PA], and [LAT]. Missing views are handled by applying a learned masking mechanism that zeroes out the corresponding projected features while retaining positional alignment. The question is prefixed with a dedicated [Q] token, and the FLAN-T5 decoder generates the answer autoregressively starting from the model's native `<BOS>` token.

This design allows the language model to jointly reason over structured multi-view representations while maintaining explicit separation between views. As a result, the model can exploit geometric and anatomical complementarities across radiographic projections, leading to improved medical visual question answering performance.

6. Experimentation and Results

All experiments were performed on a single NVIDIA RTX 4090 GPU with 24GB of memory, providing enough computational performance for training complex models.

For evaluation, we employed token-based metrics, BLEU (Papineni et al., 2002), METEOR (Banerjee and Lavie, 2005), ROUGE-L (Lin, 2004), and CIDEr (Vedantam et al., 2015) as performance metrics to assess the effectiveness of our model. Also, we have added BERTScore (Zhang et al., 2020) to have a score closer to a personal judgment, which token-based metrics cannot provide. The performance for each model is shown in Table 4.

Table 4: Comparison of our models' performance. The best results are shown in **bold**.

Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr	BERTScore
BLIP-2	0.1357	0.0954	0.0705	0.0523	0.1300	0.3472	0.7657	0.5301
VED (Train Decoder FT-Swin)	0.2837	0.2132	0.1669	0.1323	0.1707	0.4125	1.6421	0.5937
VED (+ Unfreeze Swin)	0.2867	0.2165	0.1700	0.1352	0.1724	0.4137	**1.6514**	0.5954
VED (+ RL)	**0.3046**	**0.2290**	**0.1790**	**0.1416**	**0.1756**	**0.4143**	1.6329	**0.5957**

The results in Table 4 show a clear contrast between the BLIP-2-MultiView baseline and our SwinVED models. The BLIP-2-MultiView model, despite its strong pretrained multimodal architecture, achieves lower scores across all metrics when evaluated directly on our Medical VQA task.

In contrast, the SwinVED pipeline exhibits consistent and substantial improvement through its multi-stage training strategy. Starting with the frozen Swin encoder fine-tuned on MIMIC-CXR from our previous work (Marhuenda et al., 2025), training only the decoder already yields strong baseline performance. In the second stage, unfreezing the swin encoder for joint optimization further enhances the model's capacity, improving all metrics. Finally, incorporating reinforcement learning (RL) in the third stage leads to the best performance overall, achieving the highest scores in BLEU, METEOR, ROUGE-L, and BERTScore.

These results demonstrate that, although BLIP-2-MultiView provides a strong pretrained multimodal baseline, our progressively optimized SwinVED-SCST approach is more effective for domain-specific Medical Visual Question Answering.

7. Conclusion

In this work, we introduced a large-scale medical VQA dataset for chest radiography containing over 3.2 million clinically grounded QA pairs generated through a controlled LLM pipeline. Our dataset addresses key limitations of prior work by providing diverse, evidence-based answers with strong linguistic and clinical fidelity. We also introduced two baseline models SwinVED-SCST and BLIP-2-MultiView which establish the first benchmark on this dataset and highlight its value for multimodal reasoning in radiology.

The large-scale, evidence-based nature of our LLM-generated QA pairs addresses the critical challenge of data scarcity and generalization in medical VQA. By producing a massive corpus that remains consistent with the original reports, the dataset offers a robust, clinically aligned foundation for training models that generalize effectively across varied presentations and linguistic patterns.

Our dataset creation methodology, which intentionally produces open-ended, non-templated responses, shifts medical VQA from a classification-like task to a genuine generative reasoning problem. This linguistic richness and focus on clinical fidelity make the dataset well suited for evaluating whether a model can not only detect findings but also synthesize clear, descriptive answers, promoting more human-readable and clinically useful AI systems.

Both baselines are designed as starting points for future research, offering accessible, extensible frameworks for developing more advanced medical VQA systems. By releasing the dataset and code openly, we aim to support progress toward clinically meaningful AI tools capable of assisting radiologists with accurate and interpretable question answering.

Acknowledgments

This work was supported by the Generalitat Valenciana under the grant CIPROM/2023/17.

References

Satanjeev Banerjee and Alon Lavie. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In Jade Goldstein, Alon Lavie, Chin-Yew Lin, and Clare Voss, editors, *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan, June 2005. Association for Computational Linguistics. URL https://aclanthology.org/W05-0909/.

Zeming Chen, Alejandro Hernández Cano, Angelika Romanou, Antoine Bonnet, Kyle Matoba, Francesco Salvi, Matteo Pagliardini, Simin Fan, Andreas Köpf, Amirkeivan Mohtashami, Alexandre Sallinen, Alireza Sakhaeirad, Vinitra Swamy, Igor Krawczuk, Deniz Bayazit, Axel Marmet, Syrielle Montariol, Mary-Anne Hartley, Martin Jaggi, and Antoine Bosselut. Meditron-70b: Scaling medical pretraining for large language models, 2023. URL https://arxiv.org/abs/2311.16079.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. Scaling instruction-finetuned language models, 2022. URL https://arxiv.org/abs/2210.11416.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Shengfeng Ye, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wanjia Zhao, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin

Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanhong Xu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning, 2025. URL https://arxiv.org/abs/2501.12948.

Joseph Fleiss. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76:378–382, 11 1971. doi: 10.1037/h0031619.

Kilem Gwet. Inter-rater reliability: Dependency on trait prevalence and marginal homogeneity. *Stat Methods Inter-Rater Reliab Assess*, 2, 01 2002.

Kilem Gwet. *Handbook of inter-rater reliability: The definitive guide to measuring the extent of agreement among raters.* 01 2012.

Xuehai He, Yichen Zhang, Luntian Mou, Eric Xing, and Pengtao Xie. Pathvqa: 30000+ questions for medical visual question answering, 2020. URL https://arxiv.org/abs/2003.10286.

Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, Jayne Seekins, David A. Mong, Safwan S. Halabi, Jesse K. Sandberg, Ricky Jones, David B. Larson, Curtis P. Langlotz, Bhavik N. Patel, Matthew P. Lungren, and Andrew Y. Ng. Chexpert: a large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'19/IAAI'19/EAAI'19. AAAI Press, 2019. ISBN 978-1-57735-809-1. doi: 10.1609/aaai.v33i01.3301590. URL https://doi.org/10.1609/aaai.v33i01.3301590.

Alistair E. W. Johnson, Tom J. Pollard, Seth J. Berkowitz, Nathaniel R. Greenbaum, Matthew P. Lungren, Chih-ying Deng, Roger G. Mark, and Steven Horng. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific Data*, 6(1):317, 2019. ISSN 2052-4463. doi: 10.1038/s41597-019-0322-0. URL https://doi.org/10.1038/s41597-019-0322-0.

J. Richard Landis and Gary G. Koch. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174, March 1977. ISSN 0006-341X.

Jason J. Lau, Soumya Gayen, Asma Ben Abacha, and Dina Demner-Fushman. A dataset of clinically generated visual questions and answers about radiology images. *Scientific Data*, 5(1):180251, 2018. ISSN 2052-4463. doi: 10.1038/sdata.2018.251. URL https://doi.org/10.1038/sdata.2018.251.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models, 2023. URL https://arxiv.org/abs/2301.12597.

Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics. URL https://aclanthology.org/W04-1013/.

Bo Liu, Li-Ming Zhan, Li Xu, Lin Ma, Yan Yang, and Xiao-Ming Wu. Slake: A semantically-labeled knowledge-enhanced dataset for medical visual question answering. In *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pages 1650–1654, 2021a. doi: 10.1109/ISBI48211.2021.9434010.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021b.

Zilin Lu, Yutong Xie, Qingjie Zeng, Mengkang Lu, Qi Wu, and Yong Xia. Spot the Difference: Difference Visual Question Answering with Residual Alignment . In *proceedings of Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, volume LNCS 15005. Springer Nature Switzerland, October 2024.

Luis-Jesus Marhuenda, Miquel Obrador-Reina, Mohamed Aas-Alas, Alberto Albiol, and Roberto Paredes. Unveiling differences: A vision encoder-decoder model for difference medical visual question answering. In *Medical Imaging with Deep Learning*, 2025. URL https://openreview.net/forum?id=8CNssOg7fk.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, page 311–318, USA, 2002. Association for Computational Linguistics. doi: 10.3115/1073083.1073135. URL https://doi.org/10.3115/1073083.1073135.

Steven J. Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. Self-critical sequence training for image captioning. *CoRR*, abs/1612.00563, 2016. URL http://arxiv.org/abs/1612.00563.

Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. Eva-clip: Improved training techniques for clip at scale, 2023. URL https://arxiv.org/abs/2303.15389.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien

Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models, 2023. URL https://arxiv.org/abs/2302.13971.

Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation, 2015. URL https://arxiv.org/abs/1411.5726.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert, 2020. URL https://arxiv.org/abs/1904.09675.

Appendix A. Question List

- **Atelectasis:** Is Atelectasis observed? Can Atelectasis be identified? Are there signs of Atelectasis? What evidence of Atelectasis is present? Does the patient have Atelectasis? Are there indications of Atelectasis?

- **Cardiomegaly:** Is Cardiomegaly present? Can Cardiomegaly be detected? Are there signs of Cardiomegaly? What suggests Cardiomegaly in the findings? Is there evidence of Cardiomegaly? Does the image show Cardiomegaly?

- **Consolidation:** Is Consolidation observed? Can Consolidation be identified? Are there signs of Consolidation? What evidence of Consolidation is visible? Does the image indicate Consolidation? Are there indications of Consolidation?

- **Edema:** Is Edema observed? Can Edema be detected? Are there signs of Edema? What evidence of Edema is present? Does the image suggest Edema? Are there indications of Edema?

- **Enlarged Cardiomediastinum:** Is the Cardiomedistinum enlarged? Can an enlarged Cardiomedistinum be detected? Are there signs of an enlarged Cardiomedistinum? What suggests an enlarged Cardiomedistinum? Is the Cardiomedistinum abnormal in size? Are there indications of an enlarged Cardiomedistinum?

- **Fracture:** Is a Fracture visible? Can a Fracture be identified? Are there signs of a Fracture? What evidence of a Fracture is present? Does the image indicate a Fracture? Are there indications of a Fracture?

- **Lung Lesion:** Is a Lung Lesion observed? Can a Lung Lesion be detected? Are there signs of a Lung Lesion? What suggests a Lung Lesion in the findings? Is there evidence of a Lung Lesion? Does the image show a Lung Lesion?

- **Lung Opacity:** Is Lung Opacity present? Can Lung Opacity be identified? Are there signs of Lung Opacity? What evidence of Lung Opacity is visible? Does the image indicate Lung Opacity? Are there indications of Lung Opacity?

- **No Finding:** Are there any findings? Is there any finding in the image? Can any abnormalities be detected? Are there indications of findings in the image? Are there any abnormalities present? Do you see any findings in the image?

- **Pleural Effusion:** Is Pleural Effusion present? Can Pleural Effusion be detected? Are there signs of Pleural Effusion? What evidence of Pleural Effusion is visible? Does the image suggest Pleural Effusion? Are there indications of Pleural Effusion?
- **Pleural Other:** Is there any Pleural abnormality? Can Pleural findings be identified? Are there signs of Pleural conditions? What Pleural abnormalities are present? Are there unusual Pleural findings? Does the image suggest any Pleural conditions?
- **Pneumonia:** Is Pneumonia observed? Can Pneumonia be detected? Are there signs of Pneumonia? What evidence of Pneumonia is present? Does the image indicate Pneumonia? Are there indications of Pneumonia?
- **Pneumothorax:** Is Pneumothorax present? Can Pneumothorax be detected? Are there signs of Pneumothorax? What evidence of Pneumothorax is visible? Does the image suggest Pneumothorax? Are there indications of Pneumothorax?
- **Support Devices:** Which Support Devices are found? What types of support devices are present? Which support devices can be identified? Are there any support devices observed? What support devices are detected? Can any support devices be found?
- **Heart:** Is the heart size normal? Does the heart appear enlarged? How is the size of the heart? Is the heart size within the expected range? Does the radiograph suggest a normal heart size? Is there any indication of an enlarged heart?

Appendix B. System Prompt

This appendix presents the complete system prompt used to guide the model's behavior during dataset generation

System Prompt

Task Definition: You are an assistant that, given a medical report about a chest radiograph, an indication for the imaging study and a related question, will return a JSON response with the question and an answer strictly based on the radiographic findings.

Response Format: Your response must follow this format: `{"question": "question", "answer": "Your generated answer here."}`

Strict Evidence-based Approach: The answer must include all relevant details present in the report without making inferences or assumptions beyond what is visible in the images. Ensure your answer fully addresses the question and is complete.

Prohibited Terms and Constraints: Do not use words that imply comparison or change, such as *unchanged, new, previous, compared, remains, alteration, progressive, stable, worsening, improving, evolving, resolving, persistent, ongoing, recurrent, interval, follow-up.*

Terminology Guidelines: Do not reference reports directly; instead, refer to them as images or radiographs (e.g., ”as seen in the radiograph” instead of ”as mentioned in the report”). Avoid verbs like *mentioned* or *described* and instead use direct observational terms such as *seen, show, observe, identify, demonstrate.*

Neutrality and Objectivity: Maintain an objective and neutral tone, strictly reporting what is present in the radiographic findings. Ensure that the response describes the findings as if the radiograph is being observed for the first time, avoiding terms that imply prior knowledge, comparison, or assumptions about previous states. Rephrase statements to remove any relative assessments (e.g., replace ”relative increase in opacity” with ”opacity is observed”).
figureSystem prompt.

Appendix C. User Prompt

This appendix provides the template used for the user-side input during dataset creation.

User Prompt

```
indication: {indication}
report findings: {findings}
question: {question}
```